Predicting Poverty levels in South Africa

Jessica van der Berg[a]

[a] *Stellenbosch University, South Africa*

Abstract

Abstract to be written here. The abstract should not be too long and should provide the reader with a good understanding what you are writing about. Academic papers are not like novels where you keep the reader in suspense. To be effective in getting others to read your paper, be as open and concise about your findings here as possible. Ideally, upon reading your abstract, the reader should feel he / she must read your paper in entirety.

Keywords: Multivariate GARCH, Kalman Filter, Copula

JEL classification L250, L100

1. Introduction

Since apartheid ended in 1994, the South African Government has committed a significant number of resources to distribute grants effectively and efficiently to the poor and the vulnerable. However, South Africa has remained a country with extremely high levels of inequality and poverty for an upper middle-income country Stats (2011). Previous literature has shown that poverty levels are higher in rural areas than in urban areas. This is largely due to rural household not having access to the same employment opportunities has urban households. South Africa is a country with high levels of unemployment and extremely low wages. This implies that individuals that are economically active can still fall below the upper or lower bound poverty line due to the low wages that they receive Leibbrandt, Woolard, Finn & Argent (2010).

Poverty remains unacceptably high for a country of South Africa's economic status and remains closely associated with race. Thus, poverty reduction remains one of the key economic goals. Poverty in money terms has declined markedly since apartheid ended in 1994. This was made possible by the expansion of the social grant system. However, accurately targeting social welfare programs can be challenging given that income data is often incorrect Yu & Van der Berg (2017). To overcome this problem, households are subject to a proxy means test (PMT) to identify whether a household qualifies for social assistant. This essay will try to identify new methods to identify households that

Email address: `20190565@sun.ac.za` (Jessica van der Berg)

are below the food poverty line, lower-bound poverty line and upper bound poverty line using machine learning techniques.

Structure . . . ‘

2. Literature Review

Write a short literature review here on the background of SOuth Africa - max one page

3. Data

The data used for predicting poverty level was exacted from the General Household Survey (GHS) 2018, which is a survey completed annually by Statistics SA to measure the living circumstances of households in South Africa. They survey includes household and individual characteristics. After taking out all of the NA values and taking out households who did not know or answer the relevant questions, the dataset consists of 14 546 entries.

Since the GHS consist of household survey, total income per household is reported. To calculate the average monthly income per individual within each household, I first need to calculate the total number of adults within each household. This is done by taking the difference between household size and the number of children under the age of 17. I then take the total monthly income and divide it by the total number of adults to get the average monthly income per individual in each household. This income information is then used to divided the data into four groups.

The first group is individuals whose income fall below the food poverty line. In 2018, the food poverty line was R547 per month. The food poverty line is also referred to as the extreme poverty line as it refers to the absolute minimum amount an individual will need to be able to afford the minimum energy intake for survival. The second group consist of individuals whose monthly income falls between the food poverty line and the lower-bound poverty line (R785). The lower bound poverty line is the sum of the food poverty line and and minimum amount for non-food items. The third group consist of individuals between the lower-bound poverty line and the upper-bound poverty line (R1183). The final group consist of individuals whose monthly income is above the upper-bound poverty line, which I refer to as non-vulnerable individuals. Figure 1 graphically displays the process described above in the format of a decision tree, where 1 represents the food poverty line, 2 the lower-bound poverty line, 3 the upper-bound poverty line and 4 the non-vulnerable.

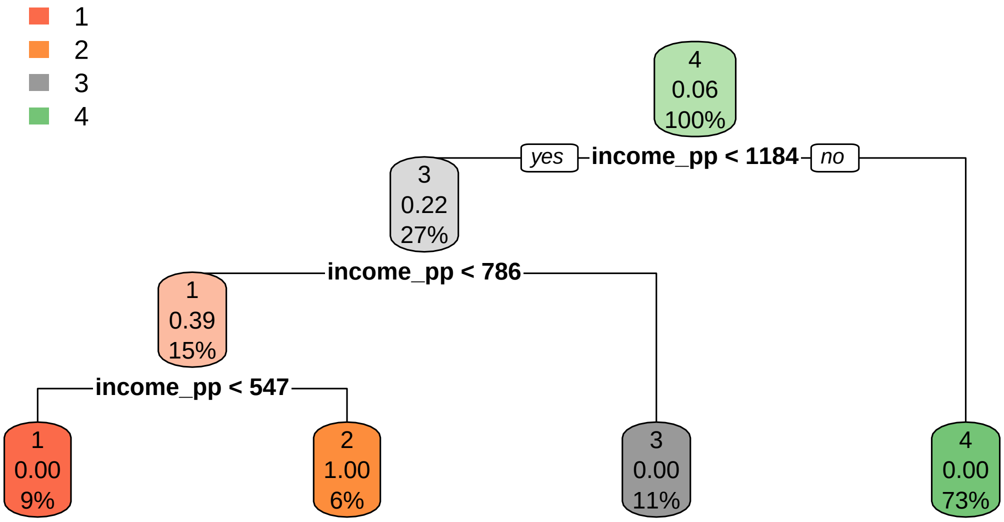

Figure 3.1: Decision tree for poverty levels

After all data cleaning was done, 73 percent of households were non-vulnerable, 11 percent were between the upper and lower-bound poverty line, 6 percent between the lower and food poverty line, and 9 percent fell below the food poverty line. To evaluate the performance for the machine learning techniques that I will implement, I randomly split the data into two subsets using a 70:30 ratio. The training dataset will consist of 70 percent of the original dataset, while the test dataset will consist of the remaining 30 percent. Figure 2 below shows the number of household per poverty level for the training dataset.

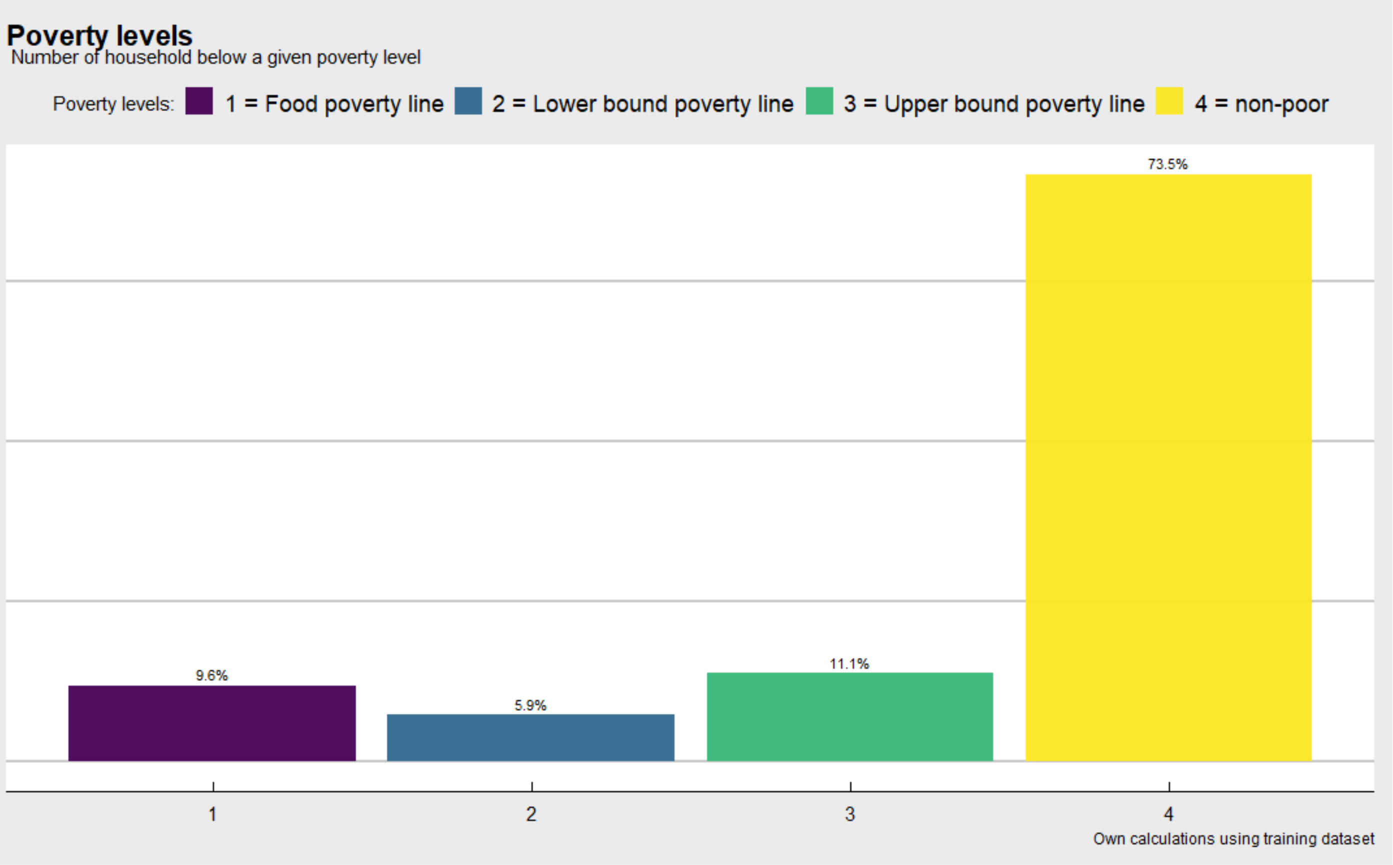

Figure 3.2: Number of household per poverty level

Figure 2 shows that in our training dataset, 73.5 percent of households are non-vulnerable and 9.6 percent of households monthly income falls below the food poverty line. This implies that I have an extremely imbalanced dataset. This is important as it will have an effect on the machine learning techniques that I implement later on.

For some descriptive statistics, I analyze some of the variable in more details. From the density plots in figure 3.3, I can see that households whose average income per person falls below the upper-bound poverty line (green line) tend to have slightly older head of household (head_age) then those household belonging to other poverty levels.

A household size (hholdsz) of 1-2 individuals tend have a larger probability to be non-vulnerable (yellow line), where larger household size of 3-5 individuals then have a larger probability of falling below the food poverty line.

Total expenditure (Q814Exp) is higher for the non-vulnerable households, which makes sense as they receive a higher income.

The total number of rooms (Q55TotRm) seem to be distributed similarly, with non-vulnerable households having a slightly smaller distribution. This comes as no suprise as these non-vulnerable households tend to have a smaller household size, implying that they need less bedrooms.

Non-vulnerable household also have much larger property valuation (Q58Val), which could imply that they have a higher standard of living.

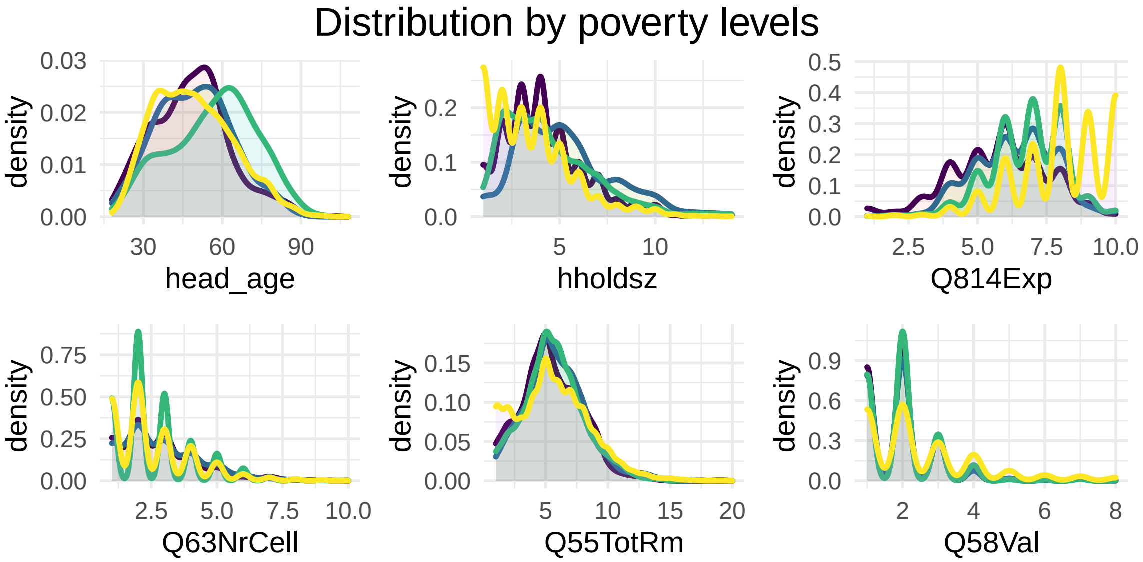

Figure 3.3: Distribution of certain variables by poverty level

Before we start building models we can also analysis which variables are correlated with our dependent variable (which is is poverty level). Figure 3.4 displays the top 10 variables that are correlated with the dependent variable. Our top four that are positively correlated are *econact hh*, which is a binary variable indicating whether or not the household is economically active, *Q814Exp*, which is a numeric variable display total household expenditure, *totmhinc*, which is a numeric variable displaying total household income, and *Q89aGrant*, which is a binary variable indicating whether the household receives a grant or not.

variables that are negatively correlated with the dependent variable is the sex of the head of the household (*head sex*), whether or not the household owns a washing machine (*Q821WashM*), the household size (*hholdsz*), the amount spent on rent or mortgage of property per month (*Q57Rent*) and the number of adults in the household (*adults*).

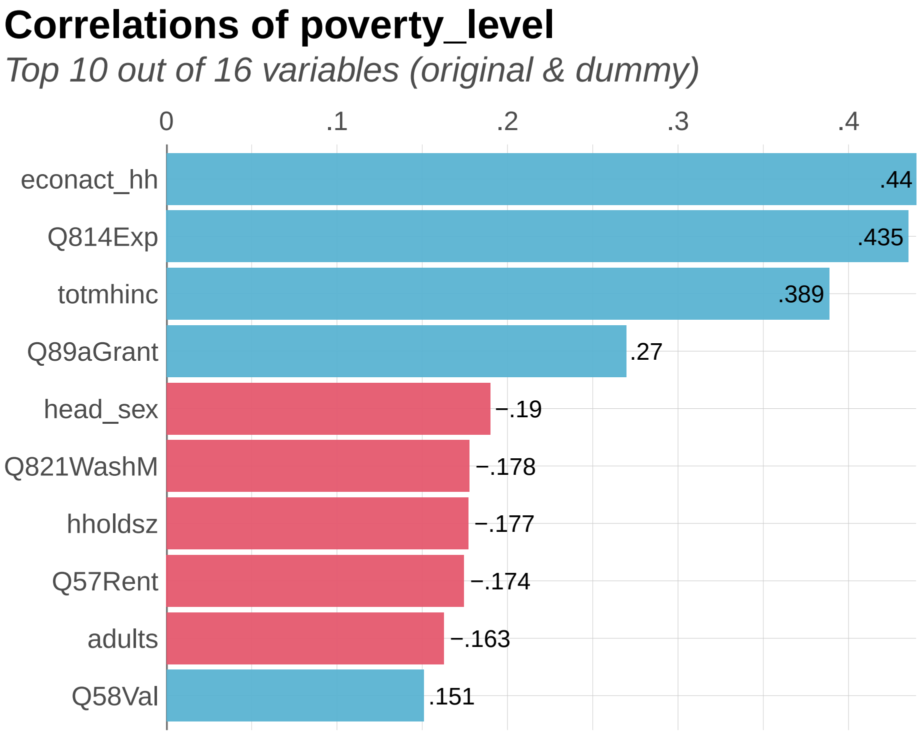

Figure 3.4: Correlation with dependent variable

4. Methodology

4.1. Subsection

Ideally do not overuse subsections. It equates to bad writing.[1]

4.2. Math section

Equations should be written as such:

$$\begin{gathered}\beta = \sum_{i=1}^{\infty} \frac{\alpha^2}{\sigma^2_{t-1}} \\ \int_{x=1}^{\infty} x_i = 1\end{gathered} \tag{4.1}$$

[1] This is an example of a footnote by the way. Something that should also not be overused.

If you would like to see the equations as you type in Rmarkdown, use $ symbols instead (see this for yourself by adjusted the equation):

$$
\beta = \sum_{i=1}^{\infty}\frac{\alpha^2}{\sigma_{t-1}^2}\int_{x=1}^{\infty}x_i = 1
$$

Note again the reference to equation 4.1. Writing nice math requires practice. Note I used a forward slashes to make a space in the equations. I can also align equations using **&**, and set to numbering only the first line. Now I will have to type “begin equation’ ’ which is a native LaTeXcommand. Here follows a more complicated equation:

$$
\begin{aligned}
y_t &= c + B(L)y_{t-1} + e_t \\
e_t &= H_t^{1/2} z_t; \quad z_t \sim N(0, I_N) \quad \& \quad H_t = D_t R_t D_t \\
D_t^2 &= \sigma_{1,t}, \dots, \sigma_{N,t} \\
\sigma_{i,t}^2 &= \gamma_i + \kappa_{i,t} v_{i,t-1}^2 + \eta_i \sigma_{i,t-1}^2, \quad \forall i \\
R_{t,i,j} &= diag({Q_{t,i,j}}^{-1}).Q_{t,i,j}.diag(Q_{t,i,j}^{-1}) \\
Q_{t,i,j} &= (1 - \alpha - \beta)\bar{Q} + \alpha z_t z_t' + \beta Q_{t,i,j}
\end{aligned} \tag{4.2}
$$

Note that in 4.2 I have aligned the equations by the equal signs. I also want only one tag, and I create spaces using “quads’ ’.

See if you can figure out how to do complex math using the two examples provided in 4.1 and 4.2.

5. Results

Tables can be included as follows. Use the *xtable* (or kable) package for tables. Table placement = H implies Latex tries to place the table Here, and not on a new page (there are, however, very many ways to skin this cat. Luckily there are many forums online!).

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
2	21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
3	22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
4	21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
5	18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00

Table 5.1: Short Table Example

To reference calculations **in text**, *do this:* From table 5.1 we see the average value of mpg is 20.98.

Including tables that span across pages, use the following (note that I add below the table: “continue on the next page’ ’). This is a neat way of splitting your table across a page.

Use the following default settings to build your own possibly long tables. Note that the following will fit on one page if it can, but cleanly spreads over multiple pages:

Table 5.2: Long Table Example

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
	21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
	21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
	22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
	21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
	18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
	18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00
	14.30	8.00	360.00	245.00	3.21	3.57	15.84	0.00	0.00	3.00	4.00
	24.40	4.00	146.70	62.00	3.69	3.19	20.00	1.00	0.00	4.00	2.00
	22.80	4.00	140.80	95.00	3.92	3.15	22.90	1.00	0.00	4.00	2.00
	19.20	6.00	167.60	123.00	3.92	3.44	18.30	1.00	0.00	4.00	4.00
	17.80	6.00	167.60	123.00	3.92	3.44	18.90	1.00	0.00	4.00	4.00
	16.40	8.00	275.80	180.00	3.07	4.07	17.40	0.00	0.00	3.00	3.00
	17.30	8.00	275.80	180.00	3.07	3.73	17.60	0.00	0.00	3.00	3.00
	15.20	8.00	275.80	180.00	3.07	3.78	18.00	0.00	0.00	3.00	3.00
	10.40	8.00	472.00	205.00	2.93	5.25	17.98	0.00	0.00	3.00	4.00
	10.40	8.00	460.00	215.00	3.00	5.42	17.82	0.00	0.00	3.00	4.00
	14.70	8.00	440.00	230.00	3.23	5.34	17.42	0.00	0.00	3.00	4.00
	32.40	4.00	78.70	66.00	4.08	2.20	19.47	1.00	1.00	4.00	1.00
	30.40	4.00	75.70	52.00	4.93	1.61	18.52	1.00	1.00	4.00	2.00
	33.90	4.00	71.10	65.00	4.22	1.83	19.90	1.00	1.00	4.00	1.00

Continued on next page

Table 5.2: Long Table Example

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.50	4.00	120.10	97.00	3.70	2.46	20.01	1.00	0.00	3.00	1.00
15.50	8.00	318.00	150.00	2.76	3.52	16.87	0.00	0.00	3.00	2.00
15.20	8.00	304.00	150.00	3.15	3.44	17.30	0.00	0.00	3.00	2.00
13.30	8.00	350.00	245.00	3.73	3.84	15.41	0.00	0.00	3.00	4.00
19.20	8.00	400.00	175.00	3.08	3.85	17.05	0.00	0.00	3.00	2.00
27.30	4.00	79.00	66.00	4.08	1.94	18.90	1.00	1.00	4.00	1.00
26.00	4.00	120.30	91.00	4.43	2.14	16.70	0.00	1.00	5.00	2.00
30.40	4.00	95.10	113.00	3.77	1.51	16.90	1.00	1.00	5.00	2.00
15.80	8.00	351.00	264.00	4.22	3.17	14.50	0.00	1.00	5.00	4.00
19.70	6.00	145.00	175.00	3.62	2.77	15.50	0.00	1.00	5.00	6.00
15.00	8.00	301.00	335.00	3.54	3.57	14.60	0.00	1.00	5.00	8.00
21.40	4.00	121.00	109.00	4.11	2.78	18.60	1.00	1.00	4.00	2.00

5.1. Huxtable

Huxtable is a very nice package for making working with tables between Rmarkdown and Tex easier.

This cost some adjustment to the Tex templates to make it work, but it now works nicely.

See documentation for this package here. A particularly nice addition of this package is for making the printing of regression results a joy (see here). Here follows an example:

If you are eager to use huxtable, comment out the Huxtable table in the Rmd template, and uncomment the colortbl package in your Rmd's root.

Note that I do not include this in the ordinary template, as some latex users have complained it breaks when they build their Rmds (especially those using tidytex - I don't have this problem as I have the full Miktex installed on mine). Up to you, but I strongly recommend installing the package manually and using huxtable. To make this work, uncomment the *Adding additional latex packages* part in yaml at the top of the Rmd file. Then comment out the huxtable example in the template below this line. Reknit, and enjoy.

Table 5.3: Regression Output

	Reg1	Reg2	Reg3
(Intercept)	-2256.361 ***	5763.668 ***	4045.333 ***
	(13.055)	(740.556)	(286.205)
carat	7756.426 ***		7765.141 ***
	(14.067)		(14.009)
depth		-29.650 *	-102.165 ***
		(11.990)	(4.635)
N	53940	53940	53940
R2	0.849	0.000	0.851

*** p < 0.001; ** p < 0.01; * p < 0.05.

FYI - R also recently introduced the gt package, which is worthwhile exploring too.

6. Lists

To add lists, simply using the following notation

- This is really simple
    - Just note the spaces here - writing in R you have to sometimes be pedantic about spaces...
- Note that Rmarkdown notation removes the pain of defining LaTeXenvironments!

7. Conclusion

I hope you find this template useful. Remember, stackoverflow is your friend - use it to find answers to questions. Feel free to write me a mail if you have any questions regarding the use of this package. To cite this package, simply type citation("Texevier") in Rstudio to get the citation for (**Texevier?**) (Note that uncited references in your bibtex file will not be included in References).

References

10 Leibbrandt, M., Woolard, I., Finn, A. & Argent, J. 2010. Trends in south african income distribution and poverty since the fall of apartheid.

Stats, S. 2011. Social profile of vulnerable groups in south africa 2002-2010. *Pretoria: Government Printer.*

Yu, D. & Van der Berg, S. 2017. South african poverty: The current situation and trends since the transition to democracy.

Appendix

Appendix A

Some appendix information here

Appendix B